Contents

Part I Machine Learning, NLP, and Speech Introduction

1 Introduction 3
- 1.1 Machine Learning 5
  - 1.1.1 Supervised Learning 5
  - 1.1.2 Unsupervised Learning 6
  - 1.1.3 Semi-Supervised Learning and Active Learning 7
  - 1.1.4 Transfer Learning and Multitask Learning 7
  - 1.1.5 Reinforcement Learning 7
- 1.2 History 7
  - 1.2.1 Deep Learning: A Brief History 8
  - 1.2.2 Natural Language Processing: A Brief History 11
  - 1.2.3 Automatic Speech Recognition: A Brief History 15
- 1.3 Tools, Libraries, Datasets, and Resources for the Practitioners 18
  - 1.3.1 Deep Learning 18
  - 1.3.2 Natural Language Processing 19
  - 1.3.3 Speech Recognition 20
  - 1.3.4 Books 21
  - 1.3.5 Online Courses and Resources 21
  - 1.3.6 Datasets 22
- 1.4 Case Studies and Implementation Details 25

References 27

2 Basics of Machine Learning 39
- 2.1 Introduction 39
- 2.2 Supervised Learning: Framework and Formal Definitions 40
  - 2.2.1 Input Space and Samples 40
  - 2.2.2 Target Function and Labels 41
  - 2.2.3 Training and Prediction 41
- 2.3 The Learning Process 42
- 2.4 Machine Learning Theory 43

2.4.1 Generalization–Approximation Trade-Off via the Vapnik–Chervonenkis Analysis 43
2.4.2 Generalization–Approximation Trade-off via the Bias–Variance Analysis 46
2.4.3 Model Performance and Evaluation Metrics 47
2.4.4 Model Validation 50
2.4.5 Model Estimation and Comparisons 53
2.4.6 Practical Tips for Machine Learning 54
2.5 Linear Algorithms 55
2.5.1 Linear Regression 55
2.5.2 Perceptron 58
2.5.3 Regularization 59
2.5.4 Logistic Regression 61
2.5.5 Generative Classifiers 64
2.5.6 Practical Tips for Linear Algorithms 66
2.6 Non-linear Algorithms 67
2.6.1 Support Vector Machines 68
2.6.2 Other Non-linear Algorithms 69
2.7 Feature Transformation, Selection, and Dimensionality Reduction 69
2.7.1 Feature Transformation 70
2.7.2 Feature Selection and Reduction 71
2.8 Sequence Data and Modeling 72
2.8.1 Discrete Time Markov Chains 72
2.8.2 Discriminative Approach: Hidden Markov Models 73
2.8.3 Generative Approach: Conditional Random Fields 75
2.9 Case Study 78
2.9.1 Software Tools and Libraries 78
2.9.2 Exploratory Data Analysis (EDA) 78
2.9.3 Model Training and Hyperparameter Search 79
2.9.4 Final Training and Testing Models 83
2.9.5 Exercises for Readers and Practitioners 85
References 85

3 Text and Speech Basics 87
3.1 Introduction 87
3.1.1 Computational Linguistics 87
3.1.2 Natural Language 88
3.1.3 Model of Language 89
3.2 Morphological Analysis 90
3.2.1 Stemming 91
3.2.2 Lemmatization 92

3.3 Lexical Representations 92
3.3.1 Tokens 92
3.3.2 Stop Words 93
3.3.3 N-Grams 93
3.3.4 Documents 94
3.4 Syntactic Representations 96
3.4.1 Part-of-Speech 97
3.4.2 Dependency Parsing 99
3.5 Semantic Representations 101
3.5.1 Named Entity Recognition 102
3.5.2 Relation Extraction 103
3.5.3 Event Extraction 104
3.5.4 Semantic Role Labeling 104
3.6 Discourse Representations 105
3.6.1 Cohesion 105
3.6.2 Coherence 105
3.6.3 Anaphora/Cataphora 105
3.6.4 Local and Global Coreference 106
3.7 Language Models 106
3.7.1 N-Gram Model 107
3.7.2 Laplace Smoothing 107
3.7.3 Out-of-Vocabulary 108
3.7.4 Perplexity 108
3.8 Text Classification 109
3.8.1 Machine Learning Approach 109
3.8.2 Sentiment Analysis 110
3.8.3 Entailment 112
3.9 Text Clustering 113
3.9.1 Lexical Chains 114
3.9.2 Topic Modeling 114
3.10 Machine Translation 115
3.10.1 Dictionary Based 115
3.10.2 Statistical Translation 116
3.11 Question Answering 116
3.11.1 Information Retrieval Based 117
3.11.2 Knowledge-Based QA 118
3.11.3 Automated Reasoning 118
3.12 Automatic Summarization 119
3.12.1 Extraction Based 119
3.12.2 Abstraction Based 120
3.13 Automated Speech Recognition 120
3.13.1 Acoustic Model 120
3.14 Case Study 122
3.14.1 Software Tools and Libraries 123
3.14.2 EDA 123

3.14.3 Text Clustering 126
3.14.4 Topic Modeling 129
3.14.5 Text Classification 131
3.14.6 Exercises for Readers and Practitioners 133
References 134

Part II Deep Learning Basics

4 Basics of Deep Learning 141
4.1 Introduction 141
4.2 Perceptron Algorithm Explained 143
4.2.1 Bias 143
4.2.2 Linear and Non-linear Separability 146
4.3 Multilayer Perceptron (Neural Networks) 146
4.3.1 Training an MLP 147
4.3.2 Forward Propagation 148
4.3.3 Error Computation 149
4.3.4 Backpropagation 150
4.3.5 Parameter Update 152
4.3.6 Universal Approximation Theorem 153
4.4 Deep Learning 154
4.4.1 Activation Functions 155
4.4.2 Loss Functions 161
4.4.3 Optimization Methods 162
4.5 Model Training 165
4.5.1 Early Stopping 165
4.5.2 Vanishing/Exploding Gradients 166
4.5.3 Full-Batch and Mini-Batch Gradient Decent 167
4.5.4 Regularization 167
4.5.5 Hyperparameter Selection 171
4.5.6 Data Availability and Quality 172
4.5.7 Discussion 174
4.6 Unsupervised Deep Learning 175
4.6.1 Energy-Based Models 175
4.6.2 Restricted Boltzmann Machines 176
4.6.3 Deep Belief Networks 178
4.6.4 Autoencoders 178
4.6.5 Sparse Coding 182
4.6.6 Generative Adversarial Networks 182
4.7 Framework Considerations 183
4.7.1 Layer Abstraction 184
4.7.2 Computational Graphs 185
4.7.3 Reverse-Mode Automatic Differentiation 186
4.7.4 Static Computational Graphs 186
4.7.5 Dynamic Computational Graphs 187

4.8 Case Study 187
4.8.1 Software Tools and Libraries 187
4.8.2 Exploratory Data Analysis (EDA) 188
4.8.3 Supervised Learning 189
4.8.4 Unsupervised Learning 193
4.8.5 Classifying with Unsupervised Features 196
4.8.6 Results 198
4.8.7 Exercises for Readers and Practitioners 198
References 199

5 Distributed Representations 203
5.1 Introduction 203
5.2 Distributional Semantics 203
5.2.1 Vector Space Model 203
5.2.2 Word Representations 205
5.2.3 Neural Language Models 206
5.2.4 word2vec 208
5.2.5 GloVe 219
5.2.6 Spectral Word Embeddings 221
5.2.7 Multilingual Word Embeddings 222
5.3 Limitations of Word Embeddings 222
5.3.1 Out of Vocabulary 222
5.3.2 Antonymy 223
5.3.3 Polysemy 224
5.3.4 Biased Embeddings 227
5.3.5 Other Limitations 227
5.4 Beyond Word Embeddings 227
5.4.1 Subword Embeddings 228
5.4.2 Word Vector Quantization 228
5.4.3 Sentence Embeddings 230
5.4.4 Concept Embeddings 232
5.4.5 Retrofitting with Semantic Lexicons 233
5.4.6 Gaussian Embeddings 234
5.4.7 Hyperbolic Embeddings 236
5.5 Applications 238
5.5.1 Classification 239
5.5.2 Document Clustering 239
5.5.3 Language Modeling 240
5.5.4 Text Anomaly Detection 241
5.5.5 Contextualized Embeddings 242
5.6 Case Study 243
5.6.1 Software Tools and Libraries 243
5.6.2 Exploratory Data Analysis 243
5.6.3 Learning Word Embeddings 244
5.6.4 Document Clustering 256

5.6.5 Word Sense Disambiguation . . . 257
5.6.6 Exercises for Readers and Practitioners . . . 259
References . . . 259

6 Convolutional Neural Networks . . . 263
6.1 Introduction . . . 263
6.2 Basic Building Blocks of CNN . . . 264
6.2.1 Convolution and Correlation in Linear Time-Invariant Systems . . . 264
6.2.2 Local Connectivity or Sparse Interactions . . . 265
6.2.3 Parameter Sharing . . . 266
6.2.4 Spatial Arrangement . . . 266
6.2.5 Detector Using Nonlinearity . . . 270
6.2.6 Pooling and Subsampling . . . 271
6.3 Forward and Backpropagation in CNN . . . 273
6.3.1 Gradient with Respect to Weights $\frac{\partial E}{\partial \mathbf{W}}$ . . . 274
6.3.2 Gradient with Respect to the Inputs $\frac{\partial E}{\partial \mathbf{X}}$ . . . 275
6.3.3 Max Pooling Layer . . . 276
6.4 Text Inputs and CNNs . . . 276
6.4.1 Word Embeddings and CNN . . . 277
6.4.2 Character-Based Representation and CNN . . . 280
6.5 Classic CNN Architectures . . . 281
6.5.1 LeNet-5 . . . 282
6.5.2 AlexNet . . . 283
6.5.3 VGG-16 . . . 285
6.6 Modern CNN Architectures . . . 285
6.6.1 Stacked or Hierarchical CNN . . . 286
6.6.2 Dilated CNN . . . 287
6.6.3 Inception Networks . . . 288
6.6.4 Other CNN Structures . . . 289
6.7 Applications of CNN in NLP . . . 292
6.7.1 Text Classification and Categorization . . . 293
6.7.2 Text Clustering and Topic Mining . . . 294
6.7.3 Syntactic Parsing . . . 294
6.7.4 Information Extraction . . . 294
6.7.5 Machine Translation . . . 295
6.7.6 Summarizations . . . 296
6.7.7 Question and Answers . . . 296
6.8 Fast Algorithms for Convolutions . . . 297
6.8.1 Convolution Theorem and Fast Fourier Transform . . . 297
6.8.2 Fast Filtering Algorithm . . . 297
6.9 Case Study . . . 300
6.9.1 Software Tools and Libraries . . . 300
6.9.2 Exploratory Data Analysis . . . 301
6.9.3 Data Preprocessing and Data Splits . . . 301

6.9.4 CNN Model Experiments 303
6.9.5 Understanding and Improving the Models 307
6.9.6 Exercises for Readers and Practitioners 309
6.10 Discussion 310
References 310

7 Recurrent Neural Networks 315
7.1 Introduction 315
7.2 Basic Building Blocks of RNNs 316
7.2.1 Recurrence and Memory 316
7.2.2 PyTorch Example 317
7.3 RNNs and Properties 318
7.3.1 Forward and Backpropagation in RNNs 318
7.3.2 Vanishing Gradient Problem and Regularization 323
7.4 Deep RNN Architectures 327
7.4.1 Deep RNNs 327
7.4.2 Residual LSTM 328
7.4.3 Recurrent Highway Networks 329
7.4.4 Bidirectional RNNs 329
7.4.5 SRU and Quasi-RNN 331
7.4.6 Recursive Neural Networks 331
7.5 Extensions of Recurrent Networks 333
7.5.1 Sequence-to-Sequence 334
7.5.2 Attention 335
7.5.3 Pointer Networks 336
7.5.4 Transformer Networks 337
7.6 Applications of RNNs in NLP 339
7.6.1 Text Classification 339
7.6.2 Part-of-Speech Tagging and Named Entity Recognition 340
7.6.3 Dependency Parsing 340
7.6.4 Topic Modeling and Summarization 340
7.6.5 Question Answering 341
7.6.6 Multi-Modal 341
7.6.7 Language Models 341
7.6.8 Neural Machine Translation 343
7.6.9 Prediction/Sampling Output 346
7.7 Case Study 348
7.7.1 Software Tools and Libraries 349
7.7.2 Exploratory Data Analysis 349
7.7.3 Model Training 355
7.7.4 Results 362
7.7.5 Exercises for Readers and Practitioners 363

7.8 Discussion . . . 364
7.8.1 Memorization or Generalization . . . 364
7.8.2 Future of RNNs . . . 365
References . . . 365

8 Automatic Speech Recognition . . . 369
8.1 Introduction . . . 369
8.2 Acoustic Features . . . 370
8.2.1 Speech Production . . . 370
8.2.2 Raw Waveform . . . 371
8.2.3 MFCC . . . 372
8.2.4 Other Feature Types . . . 376
8.3 Phones . . . 377
8.4 Statistical Speech Recognition . . . 379
8.4.1 Acoustic Model: $P(X|W)$ . . . 381
8.4.2 *LanguageModel* : $P(W)$ . . . 385
8.4.3 HMM Decoding . . . 386
8.5 Error Metrics . . . 387
8.6 DNN/HMM Hybrid Model . . . 388
8.7 Case Study . . . 391
8.7.1 Dataset: Common Voice . . . 392
8.7.2 Software Tools and Libraries . . . 392
8.7.3 Sphinx . . . 392
8.7.4 Kaldi . . . 396
8.7.5 Results . . . 401
8.7.6 Exercises for Readers and Practitioners . . . 402
References . . . 403

Part III Advanced Deep Learning Techniques for Text and Speech

9 Attention and Memory Augmented Networks . . . 407
9.1 Introduction . . . 407
9.2 Attention Mechanism . . . 408
9.2.1 The Need for Attention Mechanism . . . 409
9.2.2 Soft Attention . . . 410
9.2.3 Scores-Based Attention . . . 411
9.2.4 Soft vs. Hard Attention . . . 412
9.2.5 Local vs. Global Attention . . . 412
9.2.6 Self-Attention . . . 413
9.2.7 Key-Value Attention . . . 414
9.2.8 Multi-Head Self-Attention . . . 415
9.2.9 Hierarchical Attention . . . 416
9.2.10 Applications of Attention Mechanism in Text and Speech . . . 418
9.3 Memory Augmented Networks . . . 419
9.3.1 Memory Networks . . . 419

9.3.2 End-to-End Memory Networks . . . 422
9.3.3 Neural Turing Machines . . . 424
9.3.4 Differentiable Neural Computer . . . 428
9.3.5 Dynamic Memory Networks . . . 431
9.3.6 Neural Stack, Queues, and Deques . . . 434
9.3.7 Recurrent Entity Networks . . . 437
9.3.8 Applications of Memory Augmented Networks in Text and Speech . . . 440
9.4 Case Study . . . 440
9.4.1 Attention-Based NMT . . . 440
9.4.2 Exploratory Data Analysis . . . 441
9.4.3 Question and Answering . . . 450
9.4.4 Dynamic Memory Network . . . 455
9.4.5 Exercises for Readers and Practitioners . . . 459
References . . . 460

10 Transfer Learning: Scenarios, Self-Taught Learning, and Multitask Learning . . . 463
10.1 Introduction . . . 463
10.2 Transfer Learning: Definition, Scenarios, and Categorization . . . 464
10.2.1 Definition . . . 465
10.2.2 Transfer Learning Scenarios . . . 466
10.2.3 Transfer Learning Categories . . . 466
10.3 Self-Taught Learning . . . 467
10.3.1 Techniques . . . 468
10.3.2 Theory . . . 469
10.3.3 Applications in NLP . . . 470
10.3.4 Applications in Speech . . . 470
10.4 Multitask Learning . . . 471
10.4.1 Techniques . . . 471
10.4.2 Theory . . . 480
10.4.3 Applications in NLP . . . 480
10.4.4 Applications in Speech Recognition . . . 482
10.5 Case Study . . . 482
10.5.1 Software Tools and Libraries . . . 482
10.5.2 Exploratory Data Analysis . . . 483
10.5.3 Multitask Learning Experiments and Analysis . . . 484
10.5.4 Exercises for Readers and Practitioners . . . 489
References . . . 489

11 Transfer Learning: Domain Adaptation . . . 495
11.1 Introduction . . . 495
11.1.1 Techniques . . . 496
11.1.2 Theory . . . 513

11.1.3 Applications in NLP . . . 515
11.1.4 Applications in Speech Recognition . . . 516
11.2 Zero-Shot, One-Shot, and Few-Shot Learning . . . 517
11.2.1 Zero-Shot Learning . . . 517
11.2.2 One-Shot Learning . . . 520
11.2.3 Few-Shot Learning . . . 521
11.2.4 Theory . . . 522
11.2.5 Applications in NLP and Speech Recognition . . . 522
11.3 Case Study . . . 523
11.3.1 Software Tools and Libraries . . . 524
11.3.2 Exploratory Data Analysis . . . 524
11.3.3 Domain Adaptation Experiments . . . 525
11.3.4 Exercises for Readers and Practitioners . . . 530
References . . . 531

12 End-to-End Speech Recognition . . . 537
12.1 Introduction . . . 537
12.2 Connectionist Temporal Classification (CTC) . . . 538
12.2.1 End-to-End Phoneme Recognition . . . 541
12.2.2 Deep Speech . . . 541
12.2.3 Deep Speech 2 . . . 543
12.2.4 Wav2Letter . . . 544
12.2.5 Extensions of CTC . . . 545
12.3 Seq-to-Seq . . . 546
12.3.1 Early Seq-to-Seq ASR . . . 548
12.3.2 Listen, Attend, and Spell (LAS) . . . 548
12.4 Multitask Learning . . . 549
12.5 End-to-End Decoding . . . 551
12.5.1 Language Models for ASR . . . 551
12.5.2 CTC Decoding . . . 552
12.5.3 Attention Decoding . . . 555
12.5.4 Combined Language Model Training . . . 556
12.5.5 Combined CTC–Attention Decoding . . . 557
12.5.6 One-Pass Decoding . . . 558
12.6 Speech Embeddings and Unsupervised Speech Recognition . . . 559
12.6.1 Speech Embeddings . . . 559
12.6.2 Unspeech . . . 560
12.6.3 Audio Word2Vec . . . 560
12.7 Case Study . . . 561
12.7.1 Software Tools and Libraries . . . 561
12.7.2 Deep Speech 2 . . . 562
12.7.3 Language Model Training . . . 564
12.7.4 ESPnet . . . 566

12.7.5 Results . . . 570
12.7.6 Exercises for Readers and Practitioners . . . 571
References . . . 571

13 Deep Reinforcement Learning for Text and Speech . . . 575
13.1 Introduction . . . 575
13.2 RL Fundamentals . . . 575
13.2.1 Markov Decision Processes . . . 576
13.2.2 Value, Q, and Advantage Functions . . . 577
13.2.3 Bellman Equations . . . 578
13.2.4 Optimality . . . 579
13.2.5 Dynamic Programming Methods . . . 580
13.2.6 Monte Carlo . . . 582
13.2.7 Temporal Difference Learning . . . 583
13.2.8 Policy Gradient . . . 586
13.2.9 Q-Learning . . . 587
13.2.10 Actor-Critic . . . 588
13.3 Deep Reinforcement Learning Algorithms . . . 590
13.3.1 Why RL for Seq2seq . . . 590
13.3.2 Deep Policy Gradient . . . 591
13.3.3 Deep Q-Learning . . . 592
13.3.4 Deep Advantage Actor-Critic . . . 596
13.4 DRL for Text . . . 597
13.4.1 Information Extraction . . . 597
13.4.2 Text Classification . . . 601
13.4.3 Dialogue Systems . . . 602
13.4.4 Text Summarization . . . 603
13.4.5 Machine Translation . . . 605
13.5 DRL for Speech . . . 605
13.5.1 Automatic Speech Recognition . . . 606
13.5.2 Speech Enhancement and Noise Suppression . . . 606
13.6 Case Study . . . 607
13.6.1 Software Tools and Libraries . . . 607
13.6.2 Text Summarization . . . 608
13.6.3 Exploratory Data Analysis . . . 608
13.6.4 Exercises for Readers and Practitioners . . . 612
References . . . 612

Future Outlook . . . 615
End-to-End Architecture Prevalence . . . 615
Transition to AI-Centric . . . 615
Specialized Hardware . . . 616
Transition Away from Supervised Learning . . . 616
Explainable AI . . . 616
Model Development and Deployment Process . . . 617

Democratization of AI 617
NLP Trends 617
Speech Trends 618
Closing Remarks 618

Index 619